Stein's Method

Matthew Kwan

October 8, 2013

Commentary in blue

Issues in red

Contents

1 Introduction

introduce with limit theorems: Central Limit theorem, Poisson Limit theorem. Failure of limit theorems: they provide no understanding of speed of convergence, in particular convergence cannot be assumed to be uniform as parameters vary.

Stein's method is a technique for bounding the distance between distributions, with a variety of different distance metrics. Quantitative bounds can be useful in their own right, or can be further applied to prove asymptotic results.

1.1 Notation

For this thesis, the set of natural numbers $\mathbb{N}$ includes zero. We write 1_A for the characteristic function of A: $1_A(x) = 1$ if $x \in A$, otherwise $1_A(x) = 0$.

Unless otherwise specified, all asympotics are as $n \to \infty$. Apart from standard asymptotic notation, we use two notions of asymptotic equivalence: $f \sim g$ means $f = g(1 + o(1))$ and $f \asymp g$ means $f = O(g)$ and $g = O(f)$.

Part I

Theory

2 General Probability Theory

2.1 Review of basic concepts

I'm a little bit uncertain how much depth to go into for this. At the moment, it's written so that someone who's seen measure theory but no probability theory (an analyst) can understand.

Where possible, I've tried to translate things into the discrete case, because it's often more intuitive (and since I plan for applications to be combinatorial).

For many combinatorial applications, an informal understanding of probability theory will suffice. However, in this thesis a rigorous foundation in probability theory will be useful. The following is intended only as a brief review.

Definition 2.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity. An *event* is a measurable set $A \in \mathcal{A}$.

Definition 2.2. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity. An *event* is a measurable set $A \in \mathcal{A}$.

For our purposes Ω will often be a finite set of combinatorial objects, with $\mathcal{A}$ as the power set of Ω. In this case $\mathbb{P}$ is defined by $\mathbb{P}(\omega) := \mathbb{P}(\{\omega\})$, for each $\omega \in \Omega$. We will discuss specific probability spaces on combinatorial objects in Section 3, but we include a particularly useful definition here:

Definition 2.3. In a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where Ω is finite, if $\mathbb{P}(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, then we say $\mathbb{P}$ is *uniform*.

For an event A, $\mathbb{P}(A)$ is interpreted as the "probability that A occurs". For combinatorial spaces, events are usually of the form $A = \{\omega \in \Omega : P(\omega) \text{ holds}\}$, where $P(\omega)$ is some property of an object ω. For clarity, we often abuse notation slightly and write $\mathbb{P}(P(\omega) \text{ holds})$ instead of $\mathbb{P}(A)$.

Definition 2.4. A *random element* $X : \Omega_1 \to \Omega_2$ is a measurable function from a probability space $(\Omega_1, \mathcal{A}_1, \mathbb{P})$ to some measure space $(\Omega_2, \mathcal{A}_2, \mu)$. If the target measure space is $\mathbb{R}^n$ with the Borel σ-algebra and the Lebesgue measure, then we say X is a *random vector*; if $n = 1$ then X is a *random variable*. If X only takes countably many values then we say X is *discrete*.

If the underlying probability space Ω_1 is countable, then any function is measurable.

We will often be interested in the probability that a random element takes certain values, without regard to the underlying probability space.

Definition 2.5. Suppose X is a random element with target measure space $(\Omega, \mathcal{A}, \mu)$. The *distribution* (or *law*) $\mathcal{L}_X$ of X is the pushforward measure with respect to X. That is, it is a probability measure defined by $\mathcal{L}_X(A) = \mathbb{P}\big(X^{-1}(A)\big)$ for $A \subseteq \mathcal{A}$.

It is worth noting that in fact any probability measure is the distribution of some random element. To see this, note that given a probability measure $\mathbb{P} \in \mathcal{P}(\Omega)$, we can choose $X = \mathrm{id}_\Omega$ to have $\mathcal{L}_X = \mathbb{P}$. So, it is often convenient to specify random variables by their distributions, without defining an underlying probability space. We can use slightly abusive (but standard) notation like $\mathbb{P}(X > 1)$ to denote $\mathcal{L}_X(\{x : x > 1\})$. This is equal to $\mathbb{P}(\{\omega \in \Omega : X(\omega) > 1\})$ for any particular realization of X as a function on a probability space $(\Omega, \mathcal{A}, \mu)$.

Example 2.6. If X has the normal distribution with parameters μ and σ then we say $\mathcal{L}_X = \mathcal{N}(\mu, \sigma)$; this distribution is defined by $\mathcal{L}_X(B) = \frac{1}{\sigma\sqrt{2\pi}} \int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}\, x$ for Borel B.

If X is discrete, then $\mathcal{L}_X$ is just an assignment of a probability to each possible value.

Example 2.7. If X is Poisson distributed with parameter λ, we write $\mathcal{L}_X = \mathrm{Po}(\lambda)$; this is defined by $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$.

Definition 2.8. The *expected value* of a random variable X is $\mathbb{E}X = \int x \,\mathrm{d}\, \mathcal{L}_X(x)$.

For a random variable X that takes integer values, this definition is equivalent to the well-known formula $\mathbb{E}X = \sum_{x \in \mathbb{Z}} x\mathbb{P}(X = x)$.

If we fix a particular underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can also equivalently view expectation as a bounded linear functional on the space of integrable functions: $\mathbb{E}X = \int X(\omega) \,\mathrm{d}\, \mathbb{P}$. So, $\mathbb{E}$ is defined in terms of a particular underlying probability space. Sometimes we will define a new probability space $(\Omega, \mathcal{A}, \mathbb{P}')$ by changing the measure on the same underlying set. In this case we will write $\mathbb{E}_{\mathbb{P}'}$ to indicate expecation with respect to the measure $\mathbb{P}'$, to avoid ambiguity.

In fact, the expectation functional defines its underlying probability measure, because $\mathbb{E}\,1_A = \mathbb{P}(A)$. Since the distribution of a random variable is specified by a probability measure, the distribution $\mathcal{L}_X$ of a random variable X also uniquely defines an expectation functional $\mathbb{E}_X := \mathbb{E}_{\mathcal{L}_X}$.

Definition 2.9. For two collections $S, S' \subseteq \mathcal{A}_1$ of events, we say that S and S' are *independent* if $\mathbb{P}(A \cap A') = \mathbb{P}(A)\mathbb{P}(A')$ for each $A \in S$ and $A \in S'$. If $S = \{A\}$ contained a single set, then we say A itself is independent of S'.

Definition 2.10. Let $(\Omega_1, \mathcal{A}_1, \mathbb{P})$ be a probability space and $(\Omega_2, \mathcal{A}_2, \mu)$ a measure space. Let X be a random variable $\Omega_1 \to \Omega_2$, and let S be the set of all events of the form $\{\omega \in \Omega_1 : X(\omega) \in A_2\}$ for $A_2 \in \mathcal{A}_2$. If S is independent of S' then we say X itself is independent of S'.

We can analogously say that two random variables are independent, or a random variable and an event are independent, or any similar combination.

Definition 2.11. If two objects are not independent, then we say they are *dependent*.

Definition 2.12. Suppose $X : \Omega_1 \to \Omega_2$ is a random element defined on these spaces, and $A_1 \in \mathcal{A}_1$ is an event with nonzero probability. Then the *distribution of X conditioned on A_1* is denoted by $\mathcal{L}_{X|A_1}$ and defined by $\mathcal{L}_{X|A_1}(A_2) = \mathbb{P}(X \in A_2 | A_1)$ for $A_2 \in \mathcal{A}_2$. The expected value of a random variable with distribution $\mathcal{L}_{X|A_1}$ is called the *conditional expected value of X given A_1* and is denoted $\mathbb{E}[X|A_1]$.

We can also define conditional expectation with respect to another random variable. If X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, then the sets $X_2^{-1}(B)$ for Borel B comprise a sub-σ-algebra $\mathcal{A}'$ of $\mathcal{A}$. Then, $\mu : A' \mapsto \mathbb{E}[X_1\, 1_{A'}]$ is a signed measure on $\mathcal{A}'$ that is absolutely continuous with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}'$. By the Radon-Nikodym theorem there is an $\mathcal{A}'$-measurable random variable $\mathbb{E}[X_1|X_2]$ that satisfies $\mathbb{E}[X_1\, 1_{A'}] = \mathbb{E}[\mathbb{E}[X_1|X_2]\, 1_{A'}]$ for all A' in $\mathcal{A}'$. This random variable is almost uniquely defined: for any two choices of $\mathbb{E}[X_1|X_2]$, the probability that they differ is zero.

Definition 2.13. The random variable $\mathbb{E}[X_1|X_2]$ as defined above is called the *conditional expectation of X_1 with respect to X_2*. We can also view conditional expectation as a bounded linear operator between functions: we define $\mathbb{E}^{X_2}$ by $X_1 \mapsto \mathbb{E}[X_1|X_2]$.

This definition generalizes the previous definition of expectation conditioned on an event: if $\omega \in A$ and $\mathbb{P}(A) > 0$ then $\mathbb{E}[X|\,1_A](\omega) = \mathbb{E}[X|A]$.

Note that if X_2 is discrete then we do not need to invoke Radon-Nikodym. We can define $\mathbb{E}[X_1|X_2]$ by $\mathbb{E}[X_1|X_2](\omega) = \mathbb{E}[X_1|X_2 = X_2(\omega)]$ whenever $\mathbb{P}(X_2 = X_2(\omega)) > 0$.

2.2 Coupling

Given a finite collection of measure spaces $(\Omega_1, \mathcal{A}_1, \mu_1), \ldots, (\Omega_n, \mathcal{A}_n, \mu_n)$ recall the construction of the product measure space $(\Omega, \mathcal{A}, \mu) := (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i, \prod_{i=1}^n \mu_i)$. If a random element takes values in a product space then each component is measurable, and conversely if the components of a random tuple are measurable then that tuple is measurable in the product space. So, we can make the following definitions:

Definition 2.14. Given random elements $X_1, \ldots, X_n$ on the same underlying probability space, $\mathcal{L}_{X_1,\ldots,X_n} := \mathcal{L}_{(X_1,\ldots,X_n)}$ is called the *joint distribution* of $X_1, \ldots, X_n$. Conversely, given a random tuple $(X_1, \ldots, X_n)$, each $\mathcal{L}_{X_i}$ is called a *marginal distribution*.

Suppose we have two distributions of random elements $\mathcal{L}_{X_1}$ and $\mathcal{L}_{X_1}$. *Coupling* is the technique of constructing a random ordered pair (X_1, X_2) which realizes the given distributions as marginal distributions. Usually this is done by specifying the joint distribution $\mathcal{L}_{X_1,X_2}$.

The idea is that coupling creates a particular kind of dependence between X_1 and X_2 that allows us to compare the two distributions. Often, we are able to make conclusions about the distributions $\mathcal{L}_{X_i}$ which are independent of their specific realizations as random elements in the coupling.

2.3 Markov Chains

I'll need to define Markov Chains, stationary distributions and time-reversibility.

Perhaps I should talk more generally about stochastic processes, because applying exchangeable pairs to Stein's method is has connections with Ornstein-Uhlenbeck processes and also Stein's method can be applied to Poisson processes.

2.4 The Weak Topology on Probability Measures

The main purpose of this section is to motivate the metrics usually used in Stein's method: they are all legitimate topological metrics and are consistent with the topology of convergence in distribution

Definition 2.15. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say X_n *converges in distribution* to a random variable X if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions f. Alternatively, we say $\mathcal{L}_{X_n}$ converges *weakly* to $\mathcal{L}_X$, or simply $\mathcal{L}_{X_n} \to \mathcal{L}_X$. The topology on $\mathcal{P}(\mathbb{R})$ associated with this convergence is called the *weak topology* (we will see that it is indeed a topology). Convergence in distribution of random vectors is defined component-wise.

Definition 2.16. The *distribution function* F_X of a random variable X is defined by $F_X(x) = \mathbb{P}(X \leq x)$.

Theorem 2.17. *The following are equivalent.*

1. $\mathcal{L}_{X_n} \to \mathcal{L}_X$
2. $F_{X_n}(x) \to F_X(x)$ *for all* x *where* F_X *is continuous*
3. *(Lévy's continuity theorem)* $\mathbb{E}e^{itX_n} \to \mathbb{E}e^{itX}$ *for all* $t \in \mathbb{R}$.

The equivalence of Items 1 and 2 is a well-known result called the Portmanteau Theorem.

When X and each X_n are integer random variables, then Item 2 reduces to the condition that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k. This characterization is usually used to prove the Poisson limit theorem. Lévy's continuity theorem is classically used to prove the central limit theorem, but we will not discuss it in this thesis.

For combinatorial applications, convergence in distribution can also be proved by the "method of moments": if X is the only random variable with the moments $\left(\mathbb{E}X^k\right)_{k\in\mathbb{N}}$, then $\mathcal{L}_{X_n} \to \mathcal{L}_X$ if $\mathbb{E}X_n^k \to \mathbb{E}X^k$. Convergence in distribution can also sometimes be inferred from stronger forms of convergence when X and all the X_n are coupled to the same underlying space.

A disadvantage of all these approaches is that it is difficult to quantify the rate of convergence.

In functional analysis terms, note that expectation operators are bounded linear functionals on the space of real bounded continuous functions. Then, $\mathcal{L}_{X_n} \to \mathcal{L}_X$ just means that $\mathbb{E}_{X_n} \to \mathbb{E}_X$ in the weak-star topology. Although $C_b(\mathbb{R})^*$ is not metrizable, the subspace corresponding to $\mathcal{P}(\mathbb{R})$ is in fact metrizable, with a metric called the Lévy metric. For Stein's method we will be interested in some slightly stronger metrics.

Definition 2.18. For two probability measures $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{R})$ and a collection of real measurable "test" functions $\mathcal{H}$, define $d_{\mathcal{H}}$ by $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{\mathbb{P}_1} h - \mathbb{E}_{\mathbb{P}_2} h|$. For random variables X_1, X_2, we abuse notation to write $d_{\mathcal{H}}(X_1, X_2)$ instead of $d_{\mathcal{H}}(\mathcal{L}_{X_1}, \mathcal{L}_{X_2})$.

Each $d_{\mathcal{H}}$ is non-negative, symmetric and satisfies the triangle inequality.

Definition 2.19. A set of real functions $\mathcal{H}$ is a *determining class* if $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$ implies that $\mathbb{P}_1 = \mathbb{P}_2$.

To check that $d_{\mathcal{H}}$ is a metric, we only need to check that $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = 0$ implies that $\mathbb{P}_1 = \mathbb{P}_2$. That is, we need to check that $\mathcal{H}$ is a determining class.

Definition 2.20. We define some special cases of $d_{\mathcal{H}}$.

- If $\mathcal{H}_{\mathrm{K}} = \left\{1_{(-\infty,x]} : x \in \mathbb{R}\right\}$ then $d_{\mathrm{K}} := d_{\mathcal{H}_{\mathrm{K}}}$ is called the *Kolmogorov metric.*

- If $\mathcal{H}_{\mathrm{W}}$ is the set of real functions h that satisfy $|h(x_1) - h(x_2)| \leq |x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$ (that is, the set of functions with Lipschitz constant 1), then $d_{\mathrm{W}} := d_{\mathcal{H}_{\mathrm{W}}}$ is called the *Wasserstein metric*.

- If $\mathcal{H}_{\mathrm{TV}}$ is the set of functions 1_B for Borel B, $d_{\mathrm{TV}} := d_{\mathcal{H}_{\mathrm{TV}}}$ is called the *total variation metric*.

Proposition 2.21. *The Kolmogorov, Wasserstein and total variation "metrics" are actually metrics.*

Proof. We check that $\mathcal{H}_{\mathrm{K}}$, $\mathcal{H}_{\mathrm{W}}$ and $\mathcal{H}_{\mathrm{TV}}$ are determining classes. Let $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{W}}, \mathcal{H}_{\mathrm{TV}}\}$, and suppose that $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$. It suffices to prove that $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$, since the sets $(-\infty, x]$ generate the Borel σ-algebra. For $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{TV}}\}$ this is immediate, because $\mathbb{P}_i((-\infty, x]) = \mathbb{E}_{\mathbb{P}_i} 1_{(-\infty, x]}$. So, consider, $\mathcal{H} = \mathcal{H}_{\mathrm{W}}$.

For $\varepsilon > 0$ and $x \in \mathbb{R}$, let $h_{x,\varepsilon}$ be the continuous function which takes the value 1 on the set $(-\infty, x]$, takes the value 0 on the set $[x + \varepsilon, \infty)$, and is linearly interpolated in the range $[x, x + \varepsilon]$. Since $\varepsilon h_{x,\varepsilon} \in \mathcal{H}_{\mathrm{W}}$, we have $\mathbb{E}_{\mathbb{P}_1} h_{x,\varepsilon} = \mathbb{E}_{\mathbb{P}_2} h_{x,\varepsilon}$ for each $n \in \mathbb{N}$. For each $x \in \mathbb{R}$, $h_{x,1/n} \to 1_{(-\infty,x]}$ pointwise and each $h_{x,1/n} \leq 1$ so by the dominated convergence theorem, $\mathbb{E}_{\mathbb{P}_i} h_{1/n} \to \mathbb{E}_{\mathbb{P}_i} 1_{(-\infty,x]}$ for each $i \in \{1, 2\}$. We have again proved that $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$. □

Proposition 2.22. *The topologies induced by the Kolmogorov, Wasserstein and total variation metrics are each stronger than the weak topology.*

Proof. If $d_{\mathrm{K}}(X_n, X) \to 0$ or $d_{\mathrm{TV}}(X_n, X) \to 0$ then $F_{X_n} \to F_X$ uniformly, so certainly Item 2 of Theorem 2.17 holds.

Now, suppose $d_{\mathrm{K}}(X_n, X) \to 0$. Let $d_n = \sqrt{d_{\mathrm{K}}(X_n, X)}$ and recall the definition of $h_{x,\varepsilon}$ from the proof of Proposition 2.21. Since $d_n h_{x,d_n} \in \mathcal{H}_{\mathrm{K}}$ for each $n \in \mathbb{N}$, we have $\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n} \leq d_{\mathrm{K}}(X_n, X)/d_n = d_n \to 0$ uniformly for $x \in \mathbb{R}$. Now, note that $F_X(x - \varepsilon) \leq \mathbb{E}_X h_{x-\varepsilon,\varepsilon} \leq$

$F_X(x) \leq \mathbb{E}_X h_{x,\varepsilon} \leq F_X(x+\varepsilon)$ for any random variable X. If F_X is continuous at x then

$$F_{X_n}(x) - F_X(x) \leq (\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n}) + (F_X(x+d_n) - F_X(x)) \to 0$$
$$F_{X_n}(x) - F_X(x) \geq (\mathbb{E}_{X_n} h_{x-d_n,d_n} - \mathbb{E}_X h_{x-d_n,d_n}) + (F_X(x-d_n) - F_X(x)) \to 0$$

so Item 2 of Theorem 2.17 holds. □

Proposition 2.22 tells us that we can sensibly use our metrics to quantify the distance between random variables, in a way that is consistent with distributional (weak) convergence. All three metrics are relevant in their own right, but sometimes one may be easier to work with. It is sometimes possible to transfer results between metrics, though this usually results in worse constants than working directly in the desired metric.

It may be worthwhile to actually characterize the Wasserstein, Kolmogorov and Total Variation topologies. In particular, Wikipedia says that Wasserstein convergence is just weak convergence plus convergence of the first moment.

Definition 2.23. If $F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}\,x$ then f_X is called the *Lebesgue density* of X, and X is called a *continuous* random variable.

If X is a continuous random variable, then by the Radon-Nikodym chain rule $\mathbb{E}_X h = \int_{\mathbb{R}} h(x) f_X(x)\,\mathrm{d}\,x$.

Proposition 2.24. *Let X_1, X_2 be random variables.*

1. $d_{\mathrm{K}}(X_1, X_2) \leq d_{\mathrm{TV}}(X_1, X_2)$
2. *If X_2 has Lebesgue density bounded by C, then* $d_{\mathrm{K}}(X_1, X_2) \leq \sqrt{2C d_{\mathrm{W}}(X_1, X_2)}$.

Proof. (Adapted from [Ros11, Proposition 1.2]). Item 1 is immediate from the definition. Then, as in the proof of Proposition 2.22,

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,\varepsilon} - \mathbb{E}_X h_{x,\varepsilon}) + (\mathbb{E}_X h_{x,\varepsilon} - F_X(x)) \\ &\leq d_{\mathrm{W}}(X_1, X_2)/\varepsilon + \int_x^{x+\varepsilon} h_{x,\varepsilon} f_X(x)\,\mathrm{d}\,x \\ &\leq d_{\mathrm{W}}(X_1, X_2)/\varepsilon + C\varepsilon/2 \end{aligned}$$

and similarly

$$F_{X_n}(x) - F_X(x) \quad \geq \quad -d_{\mathrm{W}}(X_1, X_2)/\varepsilon - C\varepsilon/2,$$

So, we can take $\varepsilon = \sqrt{2d_{\mathrm{W}}(X_1, X_2)/C}$ to prove Item 2. □

Example 2.25. If $\mathcal{L}_{X_2} = \mathcal{N}(0,1)$ then $d_{\mathrm{K}} \leq (2/\pi)^{1/4}\sqrt{d_{\mathrm{W}}(X_1, X_2)}$.

In a combinatorial setting, many of our results are about integer random variables. The total variation metric is usually exclusively used in this case.

Proposition 2.26. *If X_1, X_2 are integer-valued random variables, then*

$$d_{\mathrm{TV}}(X_1, X_2) = \frac{1}{2}\sum_{k\in\mathbb{Z}} |\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)|.$$

Proof. For any Borel set A, let $d_A = \mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A)$, so that $d_{\mathrm{TV}}(X_1, X_2) = \sup|d_A|$. Define

$$A_< = \{k \in \mathbb{Z} : \mathbb{P}(X_1 = k) < \mathbb{P}(X_2 = k)\},$$
$$A_> = \{k \in \mathbb{Z} : \mathbb{P}(X_1 = k) > \mathbb{P}(X_2 = k)\}.$$

For any Borel A, we have

$$\begin{aligned} d_A &= \sum_{k\in\mathbb{Z}\cap A} (\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)) \\ &\leq \sum_{k\in A_>} (\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)) \\ &= d_{A_>} \end{aligned}$$

and similarly $\mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A) \geq d_{A_<}$. Since $d_{A_>} = -d_{A_<}$, we have

$$d_{\mathrm{TV}}(X_1, X_2) = (d_{A_>} - d_{A_<})/2 = \frac{1}{2}\sum_{k\in\mathbb{Z}} |\mathbb{P}(X_1 = k) - \mathbb{P}(X_2 = k)|.$$

□

3 Random Combinatorial Structures

I'll leave this section until I'm sure what applications I'll be looking at. Definitely random graph models, probably random permutations and random matrices.

4 Stein's Method in the Abstract

There are a few quite different presentations of Stein's method. One thing I'm trying to do here is to unify Stein's functional analysis approach for exchangeable pairs [Ste86] with Ross' general presentation[Ros11].

Suppose we have a potentially complicated random variable X, and we believe the distribution of X is close to a "standard" distribution $\mathcal{L}_0$. Then, Stein's method allows us to compare the operators $\mathbb{E}_X$ and $\mathbb{E}_0 := \mathbb{E}_{\mathcal{L}_0}$. This is sometimes directly useful for approximating statistics of X (for example, $\mathbb{P}(X \in A) = \mathbb{E}_X 1_A$). However, particularly for combinatorical applications, Stein's method is most often used to bound the distance $d_{\mathcal{H}}(\mathcal{L}_X, \mathcal{L}_0)$, where the metric $d_{\mathcal{H}}$ from Definition 2.18 is defined in terms of $\mathbb{E}_X$ and $\mathbb{E}_0$.

Stein's method is motivated by the idea of a characterizing operator.

Definition 4.1. Let $\mathcal{F}_0$ be a vector space and $\mathcal{X}_0$ be a vector space of measurable functions. We say a linear operator $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is a *characterizing operator* for the distribution $\mathcal{L}_0$ if $\operatorname{im} T_0 = \mathcal{X}_0 \cap \ker \mathbb{E}_0$. For convenience, where there is no ambiguity we will often implicitly restrict $\mathbb{E}_0$ to $\mathcal{X}_0$, so we can write $\operatorname{im} T_0 = \ker \mathbb{E}_0$.

The following proposition shows why T_0 is called a characterizing operator.

Proposition 4.2. *If $\mathcal{X}_0$ is a determining class and $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_X$ then $\mathcal{L}_X = \mathcal{L}_0$.*

Proof. If $h \in \mathcal{X}_0$, then $h - \mathbb{E}_0 h \in \ker \mathbb{E}_0 = \operatorname{im} T_0$ so $\mathbb{E}_X[h - \mathbb{E}_0 h] = 0$. That is, $\mathbb{E}_X h = \mathbb{E}_0 h$ for all $h \in \mathcal{X}_0$, which means $\mathcal{L}_X = \mathcal{L}_0$ by the definition of a determining class. □

Proposition 4.3. *$T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is characterizing if and only if there is a linear operator $U_0 : \mathcal{X}_0 \to \mathcal{F}_0$ such that the following two equations hold.*

$$\mathbb{E}_0 T_0 = 0_{\mathcal{F}_0}, \tag{1}$$

$$T_0 U_0 + \mathbb{E}_0 = \mathrm{id}_{\mathcal{X}_0}. \tag{2}$$

Proof. Suppose T_0 is a characterizing operator. Equation (1) is immediate. Let $\{h_i\}_{i\in\mathcal{I}}$ be a (Hamel) basis of $\mathcal{X}_0$. For each $i \in \mathcal{I}$ we have $h_i - \mathbb{E}_0 h_i \in \ker \mathbb{E}_0$ so there is some f_i (not necessarily unique) that solves $T_0 f_i = h_i - \mathbb{E}_0 h_i$. The operator U_0 can then be defined by $\sum_{i\in\mathcal{I}} a_i h_i \mapsto \sum_{i\in\mathcal{I}} a_i f_i$, satisfying (2).

there's probably a cleaner functional analysis way to prove that. Also, is U_0 bounded?

Conversely, suppose (1) holds and U_0 exists satisfying (2). For $h \in \ker \mathbb{E}_0$ we have $T_0(U_0 h) = h$ and $h \in \operatorname{im} T_0$, so $\ker \mathbb{E}_0 \subseteq \operatorname{im} T_0$. Equation (1) immediately says that $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_0$, so T_0 is a characterizing operator. □

We'll use (2) to give two important examples of characterizing operators.

Theorem 4.4. *Define $T_{\mathcal{N}}$ by $T_{\mathcal{N}} f(x) = f'(x) - xf(x)$. Let $\mathcal{X}_{\mathcal{N}}$ be the set of integer-valued functions h that satisfy $\mathbb{E}_{\mathcal{N}}|h| < \infty$ and let $\mathcal{F}_{\mathcal{N}}$ be the set of integer-valued functions f such that $\mathbb{E}_{\mathcal{N}}|T_{\mathcal{N}} f| < \infty$. Then $T_{\mathcal{N}} : \mathcal{F}_{\mathcal{N}} \to \mathcal{X}_{\mathcal{N}}$ is a characterizing operator for $\mathcal{N}(0,1)$.*

Proof. For any $f \in \mathcal{F}_0$, integration by parts gives

$$\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} f'(t) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-t^2/2} f(t)\,\mathrm{d}\,t = 0$$

so $\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} = 0$ and (1) holds. Then, define $U_{\mathcal{N}}$ by

$$U_{\mathcal{N}} h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2}\,\mathrm{d}\,t.$$

By the product rule and the fundamental theorem of calculus, for all $h \in \mathcal{X}_{\mathcal{N}}$ we have

$$T_{\mathcal{N}} U_{\mathcal{N}} h(x) = h(x) - \mathbb{E}_{\mathcal{N}} h,$$

so (2) holds and Proposition 4.3 completes the proof. □

Example 4.5. Note that $\mathcal{H}_{\mathrm{K}} \subseteq \mathcal{X}_{\mathcal{N}}$, where $\mathcal{H}$ is as defined in Definition 2.20

Theorem 4.6. *Define* $T_{\mathrm{Po}(\lambda)}$ *by* $T_{\mathrm{Po}(\lambda)} f(k) = \lambda f(k+1) - k f(k)$. *Let* $\mathcal{X}_{\mathrm{Po}(\lambda)}$ *be the set of integer-valued functions* h *that satisfy* $\mathbb{E}_{\mathrm{Po}(\lambda)}|h| < \infty$ *and let* $\mathcal{F}_{\mathrm{Po}(\lambda)}$ *be the set of integer-valued functions* f *such that* $\mathbb{E}_{\mathrm{Po}(\lambda)}\left|T_{\mathrm{Po}(\lambda)} f\right| < \infty$. *Then* $T_{\mathrm{Po}(\lambda)} : \mathcal{F}_{\mathrm{Po}(\lambda)} \to \mathcal{X}_{\mathrm{Po}(\lambda)}$ *is a characterizing operator for* $\mathrm{Po}(\lambda)$.

Proof. For any $f \in \mathcal{F}_0$, we have

$$\mathbb{E}_{\mathrm{Po}(\lambda)} T_{\mathrm{Po}(\lambda)} f = \sum_{i=0}^{\infty} \frac{\lambda^{i+1}}{i!} f(i+1) - \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} f(i) = 0$$

so $\mathbb{E}_{\mathrm{Po}(\lambda)} T_{\mathrm{Po}(\lambda)} = 0$ and (1) holds. Then, define $U_{\mathrm{Po}(\lambda)}$ by

$$U_{\mathrm{Po}(\lambda)} h(k) = \frac{(k-1)!}{\lambda^k} \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} \big(h(i) - \mathbb{E}_{\mathrm{Po}(\lambda)} h\big).$$

Substituting and simplifying gives

$$T_{\mathrm{Po}(\lambda)} U_{\mathrm{Po}(\lambda)} h(k) = h(k) - \mathbb{E}_{\mathrm{Po}(\lambda)} h,$$

so (2) holds and Proposition 4.3 completes the proof. □

Stein chose $\mathcal{X}_0 = \big\{h : \mathbb{E}\big[\mathrm{id}_{\mathbb{R}}^k |h|\big] < \infty \text{ for all } k\big\}$, for both the Poisson and Normal case. I'm not sure why, I'll revisit this after looking at exchangeable pairs.

When we restrict our attention to integer-valued random variables, $\mathcal{X}_{\mathrm{Po}}$ is a determining class, because it contains every function of the form 1_A, where A is a set of integers. So, $T_{\mathrm{Po}(\lambda)}$ is a characterizing operator in the sense of Proposition 4.2.

4.1 Exchangeable Pairs

4.2 Size-Bias Coupling

Part II

Applications

I'd like to go into a number of small examples (perhaps interspersed in the discussion of Stein's method in Part I), but I'd like to also go through a number of "big" examples. I'd like these examples to showcase

- different types of results: most applications give quantitative estimates. [Joh11] gives a non-quantitative distributional convergence result that was not previously proved using other methods. There are also results that have no connection with distribution metrics, such as the concentration inequalities in [Ros11]. In particular, the Latin rectangle example in [Ste86] is interesting in that the final result is not probabilistic.
- different types of distributions: definitely at least the Poisson and normal case, perhaps also an example of a more exotic distribution like the one in [FG12] or perturbations of Poisson/normal distributions as in [BČX07].
- different ways to apply stein's method: definitely exchangeable pairs and probably size-biasing. Maybe also Zero-bias coupling.

References

[BČX07] Andrew D Barbour, Vydas Čekanavičius, and Aihua Xia, *On stein's method and perturbations*, arXiv preprint math/0702008 (2007).

[FG12] Jason Fulman and Larry Goldstein, *Stein's method and the rank distribution of random matrices over finite fields*, arXiv preprint arXiv:1211.0504 (2012).

[Joh11] Tobias Johnson, *Exchangeable pairs, switchings, and random regular graphs*, arXiv preprint arXiv:1112.0704 (2011).

[Ros11] Nathan Ross, *Fundamentals of stein's method*, Probab. Surv **8** (2011), 210–293.

[Ste86] Charles Stein, *Approximate computation of expectations*, Lecture Notes – Monograph Series **7** (1986).